The Probability of Exceedance as a Nonparametric Person-Fit Statistic for Tests of Moderate Length

Tendeiro, Jorge N.; Meijer, Rob R.

Published in:
Applied Psychological Measurement

DOI:
10.1177/0146621613499066

Article

The Probability of Exceedance as a Nonparametric Person-Fit Statistic for Tests of Moderate Length

Applied Psychological Measurement
37(8) 653–665

DOI: 10.1177/0146621613499066
apm.sagepub.com
SAGE

Jorge N. Tendeiro[1] and Rob R. Meijer[1]

Abstract

To classify an item score pattern as not fitting a nonparametric item response theory (NIRT) model, the probability of exceedance (PE) of an observed response vector **x** can be determined as the sum of the probabilities of all response vectors that are, at most, as likely as **x**, conditional on the test's total score. Vector **x** is to be considered not fitting when its PE is smaller than a prespecified level. Although this concept is not new, it is hardly if ever applied in practice. In the present paper, the authors show how the PE of a response vector **x** can be computed in a NIRT context and how misfitting response patterns are detected using the exact distribution of PE. Results from two empirical applications are discussed. A simulation study is conducted to investigate the robustness of the PE against violation of the invariant item ordering condition. Finally, considerations over possible asymptotic distributions of PE are discussed.

Keywords

nonparametric item response theory, aberrant response behavior, probability of exceedance, person fit

In psychological and educational testing, checking the validity of individual test scores is an important element in the assessment procedure. One way to do this is to investigate the consistency of observed item scores with the probability expected under an item response theory (IRT; Embretson & Reise, 2000) model. Analyzing how well individual response vectors fit an IRT model is known as person-fit research or appropriateness measurement (e.g., Meijer & Sijtsma, 2001). The importance of person-fit research is, for example, recognized in the guidelines of the International Test Commission (2011) that recommend checking test score validity through checking unexpected response patterns. Also, several testing organizations considered implementing methods for individual test score and item score validity. For example, at Educational Testing Service, the TOEFL program already implemented quality control charts for its quarterly reviews as preemptive checks (A. von Davier, personal communication, July 12, 2012). Another example of the usefulness of checking the consistency of individual item scores was given in Meijer, Egberink, Emons, and Sijtsma (2008), who identified schoolchildren who did not understand the phrasing of many questions of a questionnaire on self-concept.

[1]University of Groningen, Netherlands

Corresponding Author:
Jorge N. Tendeiro, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands.
Email: j.n.tendeiro@rug.nl

Furthermore, Ferrando (2012) discussed the use of person-fit research to screen for idiosyncratic answering behavior and low person reliability for students filling out a Neuroticism and Extraversion personality scale. For overviews on some of the statistics and procedures available in person-fit research, see Karabatsos (2003) and Meijer and Sijtsma (1995, 2001).

Although there are many person-fit statistics available in the literature, most statistics were proposed in the context of *parametric* IRT. In many test applications, however, it can be very convenient to have person-fit statistics that do not require the estimation of parametric IRT parameters but, instead, only require *nonparametric* IRT (NIRT) item indices such as the item proportion-correct scores. For example, for almost all tests and questionnaires that are evaluated in the Dutch Rating System for Test Quality (Evers, Sijtsma, Lucassen, & Meijer, 2010), no information is available with respect to parametric IRT parameters, and this also applies for tests in other countries (Geisinger, 2012).

Meijer and Sijtsma (2001) discussed early attempts to formulate NIRT statistics (see also Emons, 2008; Meijer, 1994). When applying these statistics, however, there is a lack of research that can help a researcher to decide when to classify an item score pattern as fitting or misfitting. To classify a pattern as (mis)fitting, a distribution is needed. This distribution can be based on empirical results, exact results, asymptotic results, or simulation. For parametric IRT, there are studies that discuss these distributions (e.g., Drasgow, Levine, & McLaughlin, 1991; Magis, Raîche, & Béland, 2012; Meijer & Tendeiro, 2012; Snijders, 2001). For nonparametric approaches, there are, however, almost no studies that discuss when to classify an item score pattern as misfitting. Although one can always use some rule-of-thumb, such as 1.5 standard deviations from the mean score (see, for example, Tukey, 1977), in many situations information about the probability of the realization of a particular score pattern would help researchers to decide when to classify a pattern as misfitting.

In the present study, the authors discuss an approach to classify a score pattern as misfitting (aberrant) without assuming a parametric model. They show that for tests of moderate length, the exact distribution of the so-called *probability of exceedance* (PE) can be used to classify a score pattern as normal or aberrant. The procedures to compute the PE for individual response vectors and the associated exact distributions are explained and are partly based on earlier work by van der Flier (1982). The practical application of the PE is illustrated with two real data sets. Furthermore, the authors show that violations of the assumption of invariant item ordering (IIO; Meijer & Egberink, 2012; Sijtsma & Junker, 1996; Sijtsma, Meijer, & van der Ark, 2011) property do not seem to have a large impact on the performance of the PE statistic. Finally, the extension of the PE statistic to long tests is discussed.

A Nonparametric Approach to Person Fit

In the present study, NIRT models (Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002) are used. The main goal in NIRT is to use the scores of a group of persons on a test to *rank* the persons on an assumed latent trait θ (and not to *estimate* each person's θ-score as in parametric IRT). Let random variable X_i denote the score on item i ($i = 1, \ldots, k$, where k is the number of items in the test or questionnaire). All items considered in this paper are dichotomous; hence, item scores are either equal to 0 or 1 (to code incorrect or correct answers, respectively). The so-called *item response function* (IRF) is defined as

$$P_i(\theta) = P(X_i = 1|\theta), \tag{1}$$

where $i = 1, \ldots, k$. In NIRT, the IRFs are defined as distribution-free functions of the latent ability. The following three assumptions in NIRT are typically used (e.g., Sijtsma & Molenaar,

2002): (a) Unidimensionality: All items in the test are designed to predominantly measure the same latent trait, θ. (b) Local independence, that is, answers to different items are statistically independent conditional on θ, therefore $P(X_1 = x_1, \ldots, X_k = x_k|\theta) = \prod_{i=1}^{k} P(X_i = x_i|\theta)$. (c) Latent monotonicity of the IRFs: Each IRF is monotone nondecreasing in θ, that is, $P_i(\theta_a) \leq P_i(\theta_b)$ for all $\theta_a < \theta_b$ and $i = 1, \ldots, k$. It can be observed that these assumptions also apply to the most common parametric IRT models (e.g., the logistic models; see Embretson & Reise, 2000). In this sense, one can regard parametric IRT models as NIRT models with added constraints (namely, in the functional relationship between θ and the probability of answering the item correctly).

The NIRT model that satisfies Assumptions 1 through 3 is known as the monotone homogeneity model (MHM). When the items are dichotomous, the MHM implies a stochastic ordering of the latent trait (SOL) by means of the total score statistic $X_+ = \sum_{i=1}^{k} x_i$ (Hemker, Sijtsma, Molenaar, & Junker, 1997):

$$P(\theta > c|\mathrm{X}_+ = s) \leq P(\theta > c|\mathrm{X}_+ = t), \tag{2}$$

for any fixed value c and for any total scores $0 \leq s < t \leq k$. SOL is an important property because it justifies the common use of the total score X_+ to infer the ordering of the persons on the unobservable latent scale.

A condition that eases the interpretation of person-fit results and that the authors use to interpret the PE is that of IIO. The k items of a test satisfy the IIO assumption if

$$P_1(\theta) \leq P_2(\theta) \leq \ldots \leq P_k(\theta) \text{ for all } \theta. \tag{3}$$

In other words, IIO means that the IRFs do not intersect (Sijtsma & Molenaar, 2002). A model verifying IIO allows for an ordering of the items that is independent from θ. In practical terms, this means that the relative difficulty of the items is the same across the entire latent scale.

IIO may seem a strong assumption that is difficult to satisfy in practice for some datasets (e.g., Ligtvoet, van der Ark, te Marvelde, & Sijtsma, 2010; Sijtsma et al., 2011; Sijtsma & Junker, 1996). However, Meijer and Egberink (2012) and Meijer, Tendeiro, and Wanders (2013) found that, for clinical scales, IIO was satisfied in many cases once low discriminating items were removed from the data. Besides, Sijtsma and Meijer (2001) showed that the performance of a number of person-fit statistics was robust against violations of IIO. The authors assume IIO when they discuss the PE. Practical consequences of violating IIO with respect to the PE statistic are addressed later in this study.

The PE

Define the *proportion-correct* score (p value) of item i by $p_i = \int_\theta P_i(\theta) f(\theta) d\theta$, where $f(\theta)$ is the density of ability θ in the population. The IIO assumption allows ordering the items by their p values, say $p_1 \leq p_2 \leq \ldots \leq p_k$, and this ordering is constant across θ. The probability that random vector $\mathbf{X} = (X_1, \ldots, X_k)$ will be equal to a specific response vector $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ can then be defined by

$$P(\mathbf{x}) = p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{k} p_i^{x_i} (1 - p_i)^{1 - x_i}. \tag{4}$$

Pattern **x** deviates from the expected score pattern if too many ''easy'' items (i.e., items with large p_i) are answered incorrectly, and/or if too many ''difficult'' items (i.e., items with low p_i) are answered correctly. In this sense, **x** is considered *deviant* or *aberrant* when it does not closely match the expected score pattern that is suggested by the population's p values. The PE of an item score pattern **x**, PE(**x**), is a measure of the deviance between **x** and the expected score pattern. PE(**x**) is defined as the sum of the probabilities of all response vectors which are, at most, as likely as **x**, given the total score X_+:

$$\mathrm{PE}(\mathbf{x}) = \sum p(\mathbf{X} = \mathbf{y} | X_+), \tag{5}$$

where the summation extends to all response vectors **y** with total score X_+ verifying $P(\mathbf{y}) \leq P(\mathbf{x})$. The authors observe that because the computation of PE relies on item response vectors that were not necessarily observed in the data (specifically, response vectors *less* likely than **x**), the PE person-fit statistic does not follow the Likelihood Principle (Birnbaum, 1962; Lee, 2004).

The PE statistic requires Assumptions 1 through 3 to hold, that is, the MHM should fit the data adequately. Conditioning the probabilities on the right-hand side of Equation 5 on the total score is based on the stochastic ordering of the subjects on the θ scale. Therefore, the PE of a score pattern **x** accumulates evidence against **x** based on a subpopulation of persons with the same latent trait (recall Equation 2). Response pattern **x** is considered deviant or aberrant with respect to the population's expected score if PE(**x**) is smaller than a predefined level (e.g., .05 or .01, or some predefined percentile of the exact distribution of PE). The IIO assumption is useful for a correct interpretation of the PE because of the unique ordering of the items by their p values across θ. Violations of IIO may lead to situations where the relative difficulty of the items change for different values of θ. Assuming IIO avoids this kind of ambiguity when interpreting item scores. Therefore, it is generally useful to use a model meeting the IIO assumption when the main goal is to compare item score patterns between persons with different total scores, as is the case in person-fit analyses.

A small example illustrating the concept of the PE is provided in the Online Appendix to this paper (see Online Appendix, Part A). Part B of the Online Appendix shows the R (R Development Core Team, 2011) code that can be used to perform all the necessary computations.

It is observed that the estimation of the cutoff level for the PE statistic is dependent on the type of data to be analyzed. Factors such as the number of items in the scale, the item p values, and the total sum-score play a role in determining the exact distribution of the PE statistic. Several approaches can be conceived to determine adequate cutoff values, such as using predefined cutoff values (1%, 5%, or 10%), using percentiles of the empirical distribution, using percentiles of the exact distribution, or estimating cutoff values using bootstrapping procedures. The researcher should decide, in each situation, which approach provides the most sensible results in terms of false/true positive rates.

Simulation Study

Before the authors discuss two empirical examples, they first investigate the performance of the PE statistic in a simulation study. The goal was to have a clearer impression concerning PE's detection rates and robustness against violations of the NIRT model assumptions, under two different settings. In particular, they were interested in studying the robustness of PE against violations of IIO using simulated data. As discussed before, IIO is a useful property in the framework of person fit because it allows a unique ordering of the items across the latent scale θ. A violation of Equation 3 may lead to intervals on the θ scale where, say, item i is easier than item j and to intervals where the reverse is also possible. Such situations should be avoided when possible.

In this section, the authors present the results of a simulation study that investigated how much the PE statistic is affected by violations of IIO (and by other necessary model assumptions, to be presented shortly). Several procedures from the *mokken* R package (van der Ark, 2007, 2012) were used to check the fit of the MHM to the data as well as violations of IIO. General guidelines given by Sijtsma et al. (2011) and Meijer and Tendeiro (2012) were followed to perform the analyses. In particular, Meijer and Tendeiro discussed that before investigating person fit, it is important to first check whether the IRT model fits the data—if not, misfit is difficult to interpret. Hence, special attention is paid to model fitting prior to person-fit assessment.

Data Simulation: Normal Response Patterns

Twenty different datasets with scores of 1,000 persons on 15 items were generated using the one-parameter logistic model with item discrimination equal to 1.7 for every item; item difficulty and person ability parameters were drawn from the standard normal distribution. The number and percentages of examinees displaying aberrant behavior equaled *N.aberr* = 10, 50, and 100, corresponding to 1%, 5%, and 10% of the examinees, respectively, and the number and percentages of items answered aberrantly equaled *k.aberr* = 3, 5, and 10, corresponding to 20%, 33%, and 66% of the items, respectively. Examinees were randomly selected to display aberrant behavior according to two criteria: The latent ability should be low (more precisely, $\theta < -1$), and the total sum-score on the (*k.aberr* + 3) most difficult items of the scale should not exceed 3. Two types of aberrant behavior were mimicked in this simulation: Cheating and random responding. In the case of cheating, *k.aberr* 0s out of the (*k.aberr* + 3) most difficult items were randomly selected and changed into 1s. In the case of random responding, each of the *k.aberr* 0s out of the (*k.aberr* + 3) most difficult items were changed into 1s with probability .25. Moreover, 20 replications were simulated for each condition. This framework served as the basis for the simulation study.

The analysis was started by confirming that some necessary conditions for the MHM to hold were met for the 20 datasets generated (prior to imputing aberrant behavior). All interitem covariances were nonnegative (Sijtsma & Molenaar, 2002, Theorem 4.1), and all item-pair scalability coefficients H_{ij} satisfied $0 < H_{ij} < 1$ (Sijtsma & Molenaar, 2002, Theorem 4.3). Moreover, no violations of monotonicity were found. The Automated Item Selection Procedure (AISP; Mokken, 1971; Sijtsma & Molenaar, 2002) was used to select items that comply with the MHM (i.e., such that all interitem covariances are positive and all item scalability coefficients H_i are larger than a specified lower bound $c = .3$). All 15 items were selected by the AISP, thus assuring the scalability of the generated set of items. The overall scalability coefficients of the 20 datasets varied between $H = .42$ and $H = .50$; hence, the scales can be considered moderate with respect to its precision to order persons on the latent scale by means of the total scores (Mokken, 1971). Also, several methods available in the *mokken* package were used to look for violations of IIO: the pmatrix method (Molenaar & Sijtsma, 2000), the restscore method (Molenaar & Sijtsma, 2000), and the manifest IIO method (MIIO; Ligtvoet et al., 2010). No significant violations of IIO were found. The H^T coefficients of the 20 datasets varied between .32 and .64 ($M = .48$, $SD = .08$), and thus, the precision of the item ordering was, on average, medium (Ligtvoet et al., 2010).

Data Simulation: Aberrant Response Patterns

Next, aberrant behavior (cheating, random responding) was inputted in each 15-item data set following the procedure previously explained. The PE was then computed for each of the 1,000 examinees. A cutoff level of .10 was used as the criterion to flag item score vectors: Vectors **x**

verifying PE(**x**) < .10 were flagged as *potentially* displaying aberrant behavior. This cutoff level kept empirical Type I error rates between 1% and 3% in case of cheating and between 2% and 3% in case of random responding (more detailed information is displayed in Part C of the Online Appendix, Tables C1 and C2).

Scores on five additional items were generated using a similar procedure as before except for the discrimination parameter, which was now fixed at 1.2 for these items. The scores on the original 15 items were combined with the scores on these extra 1, 2, . . . , 5 items. The total number of items in the scale (variable *length*) was also used as a factor to explain the findings in the simulation study. It was expected that the different discrimination values used to generate the scores on the additional set of items would lead to an increasing number of significant violations of IIO (Sijtsma & Molenaar, 2002).

Results From Simulation Study: Model Violations

The effects of *N.aberr, k.aberr*, and *length* on the mean number of significant violations to IIO were analyzed using full factorial models. Results show that factor *length* had indeed the largest effect on the number of significant violations to IIO in the random responding situation. Similar results were found in the cheating situation with the exception of the pmatrix criterion. More details about the sizes of the effects found can be consulted in Part D of the Online Appendix (top panels of Tables D1 and D2).

The authors also inspected whether other NIRT model assumptions were affected by their data manipulation (concerning the imputation of aberrant behavior and the addition of one through five extra items to the data). They checked whether interitem covariances were nonnegative (Sijtsma & Molenaar, 2002, Theorem 4.1), item-pair scalability coefficients H_{ij} were between 0 and 1 (Sijtsma & Molenaar, 2002, Theorem 4.3), and IRFs displayed monotonicity. Full factorial models (main effects: *N.aberr, k.aberr*, and *length*) were fitted. It was verified that all model assumptions were significantly affected: nonnegative interitem covariances, $F(9, 1070) = 66.42$, $p < .01$, $R^2 = .36$; H_{ij} coefficients between 0 and 1, $F(9, 1070) = 150.3$, $p < .01$, $R^2 = .56$; monotonicity, $F(9, 1070) = 91.29$, $p < .01$, $R^2 = .43$. It was also verified that *N.aberr* (the proportion of subjects in the sample displaying aberrant behavior) was the factor with the largest effect on the violations of the NIRT model conditions (more details in Part D of the Online Appendix, bottom panels of Tables D1 and D2).

Summarizing, the authors concluded that adding 1 through 5 differently discriminating items to the initial set of 15 items did lead to a significant increase on the number of violations to IIO. This effect was more evident in the random responding setting than in the cheating setting. Moreover, it was verified that other NIRT model assumptions (nonnegative interitem covariances, H_{ij} coefficient between 0 and 1, and latent monotonicity) were mostly affected by the number of subjects in the sample that displayed aberrant behavior. The authors cannot stress enough how important they find it to carefully check, and report, model assumptions before attempting any kind of person-fit analyses.

Results From Simulation Study: Detection Rate PE

The next step was to analyze the detection rate of the PE under both types of aberrant behavior considered in this study. Figures 1 and 2 (for cheating and random responding, respectively) display the detection rates found using a PE threshold value of .10; the size of the effect of each experiment factor on the detection rates can be consulted in Part E of the Online Appendix. The number of items displaying aberrant behavior (*k.aberr*) had the largest effect on the detection rates in both settings. Interestingly, the detection rates were higher for a moderate value of

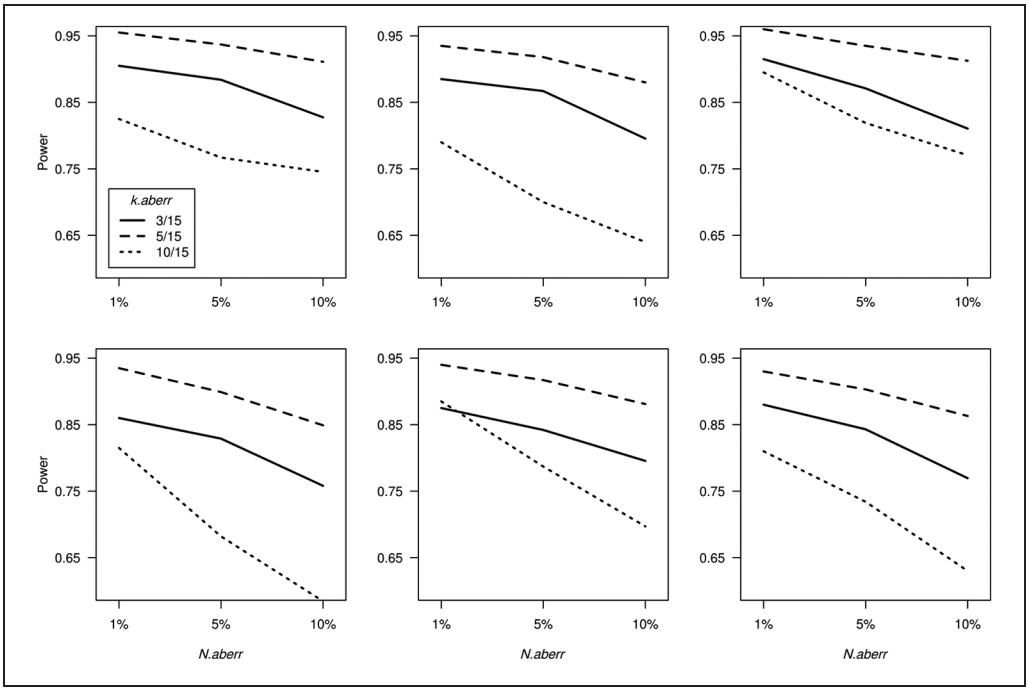

Figure 1. Cheating detection rate of the probability of exceedance for a number of items equal to 15 (top left), 16 (top middle), 17 (top right), 18 (bottom left), 19 (bottom middle), and 20 (bottom right).

k.aberr (=5); both low and large values of *k.aberr* are associated to lower power. This finding is in line with results in St-Onge, Valois, Abdous, and Germain (2011), where it was shown that the detection rates of several person-fit statistics increase with aberrance rates only to some point, after which a decrease is to be expected. It was also observed that the cheating detection rates decreased with *N.aberr*. This can be understood by observing that the cheating behavior that was imputed led to sum-score differences (before *versus* after cheating imputation), which had a large impact on the original sum-scores (which were typically very low). In other words, PE seemed to decrease its performance when the aberrance rate increased beyond moderate boundaries (St-Onge et al., 2011). The imputation of random responding behavior, on the other hand, was milder (the selected 0s were changed into 1s with probability .25). This introduced a more moderate rate of aberrant behavior in the data and the PE statistic performed accordingly (for *k.aberr* = 3, 10): Its detection rate improved with *N.aberr*. The *k.aberr* = 5 in the random responding case was different because the PE's detection rate decreased with *N.aberr*. Once more, the explanation resides in the balance that must exist between the performance of a person-fit statistic (PE in this study) and the level of aberrant rate in the data. When *k.aberr* = 5, the actual detection rate is larger than for the other values considered, but adding more and more ''aberrant'' examinees to the set did surpass some ''breakpoint'' of the PE statistic, which affected its performance for higher rates of aberrant examinees in the data.

In general, it can be concluded that the PE performed very well in the cheating case and moderately well in the random responding case. The PE statistic did not seem to be overly affected by violations of IIO or other model assumptions. Several factors, such as the number of examinees and items displaying aberrant behavior, must be taken into account when judging the performance of PE. Also, the authors stress two important ideas that should be taken into account

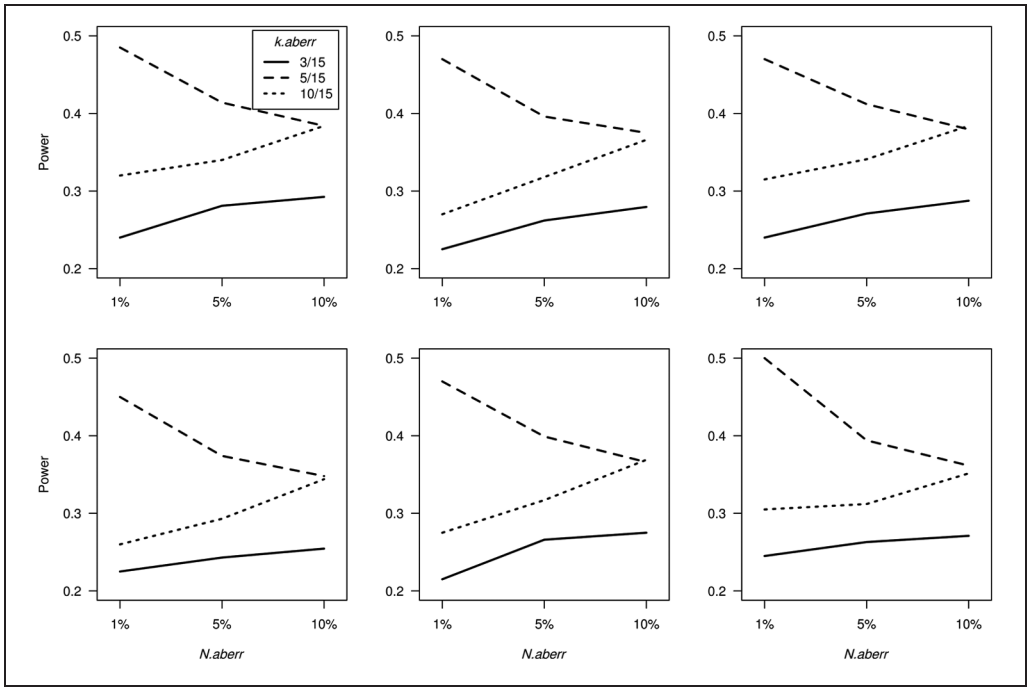

Figure 2. Random responding detection rate of the probability of exceedance for a number of items equal to 15 (top left), 16 (top middle), 17 (top right), 18 (bottom left), 19 (bottom middle), and 20 (bottom right).

when attempting to perform any person-fit analysis (using PE or any other statistic). They find it important to check whether the item response model of choice fits the data adequately (as they did) and to check how the performance of the person-fit statistic is overly affected by the several factors that play a role in fit measurement.

Two Empirical Examples

The aim of these empirical examples was twofold. On one hand, the authors would like to apply the PE method to empirical data and see whether they can interpret the results. On the other hand, they would like to expand the methodology proposed by Sijtsma et al. (2011) to check the assumptions of the MHM and the IIO condition through checking misfitting response patterns. Data of the Social Inadequacy (SI, 13 items) and the Inadequacy (IN, 28 items) subscales from the Dutch Personality Questionnaire–Junior (Dutch: Junior Nederlandse Persoonlijkheidsvragenlijst [NPV-J]; Luteijn, van Dijk, & Barelds, 2005) were analyzed. The SI and the IN scales were selected because these scales had the best psychometric properties (Weekers & Meijer, 2008). The sample consisted of scores of 866 adolescents between 9 and 15 years of age who judged themselves on the personality constructs of interest. The analysis of the SI data is reported next; the analysis of the IN data is presented in Part G of the Online Appendix.

The SI Scale

The authors confirmed that all interitem covariances were nonnegative, and all item-pair scalability coefficients H_{ij} were between 0 and 1, as required. Also, there were no significant

Table 1. Scores of the Four Subjects on the Selected Social Inadequacy Subscale of 11 Items With PE Smaller Than .01 (Top Panel), and Scores of Three Subjects With PE > .10 as Term of Comparison (Bottom Panel).

Item	si21	si89	si85	si62	si23	si22	si25	si26	si51	si79	si105	NC	PE
p values	.68	.53	.50	.45	.44	.32	.27	.25	.24	.19	.14	—	—
Person 120	0	0	0	0	0	0	0	0	1	0	1	2	.0035
Person 129	0	1	0	0	0	1	0	1	0	1	1	5	.0030
Person 310	0	0	1	0	0	0	1	0	1	1	1	5	.0013
Person 471	0	1	0	0	0	0	1	0	1	1	1	5	.0019
Person 139	0	1	1	1	1	0	1	1	0	1	0	7	.1974
Person 559	0	1	1	1	0	1	1	0	1	1	0	7	.1063
Person 832	1	0	0	1	1	0	1	0	0	0	1	5	.2124

Note. Items are ordered in increasing order of difficulty. NC = number-correct score; PE = probability of exceedance.

violations of the monotonicity assumption. However, some of the item scalability coefficients H_i reflected low item discrimination. Usually, items with H_i values below $H_i = .3$ are considered not scalable. The identified items showing low discrimination were si22, si44, and si80 ($H_i = .27$, .18, and .18, respectively). These numerical results were graphically confirmed by looking at the estimated nonparametric IRFs of all items in the SI scale (these plots are shown in Part F of the Online Appendix, Figure F1). Furthermore, the test scalability coefficient equaled $H = .34$.

Several items were involved in violations of the IIO assumption. The pmatrix, restscore, and MIIO procedures in the *mokken* package detected 9, 9, and 10 items involved in statistically significant violations of IIO, respectively. The authors therefore decided to use the AISP to search for a subset of items for which the MHM provided a more adequate fit than the full SI scale. A subscale consisting of all items of the scale except items si22, si44, and si80 was selected. However, it was decided not to exclude item si22 from the scale for two reasons: (a) This item's scalability coefficient was close to the .3 lower bound ($H_i = .27$), and (b) this item's content was not covered by other items in the scale. Thus, only items si44 and si80 were removed from the scale. The item scalability coefficients for the subscale of 11 items were all $H_i > .3$ except for item si22, but the violation was marginal ($H_i = .299$, $SD = .027$; see Table F1 in the Online Appendix). Furthermore, the test scalability coefficient equaled $H = .42$ (the updated estimated nonparametric IRFs for the items of this subscale are shown in Figure F2 in the Online Appendix). The authors checked the monotonicity assumption for this subscale using item-restscore regression and found no significant violations. Finally, the number of items involved in statistically significant violations of IIO reduced with respect to the full scale (using the pmatrix, restscore, and MIIO procedures, 2, 4, and 4 items were involved in statistically significant violations of IIO, respectively). Removing items si23 and si51 from the data would have eliminated all significant violations of IIO. However, it was decided not to remove item si23 or si51 because the IIO violations were small and did not invalidate the fit of the model (see Emons, Sijtsma, & Meijer, 2005, Sijtsma & Meijer, 2001, for robustness studies). Thus, 11 items were used to check the validity of the individual scores by means of the PE person-fit statistic.

The PE was computed using the scores of the 866 adolescents. The *p* values of these items averaged .36 ($SD = .17$). A total of 19 persons (2.2%) had a PE smaller than .05, and 4 persons (.5%) had a PE smaller than .01. For these persons, there is strong evidence that some kind of aberrant behavior occurred. Table 1 (top panel) shows the scores of the four persons whose PE

statistic was smaller than .01. Note that for these patterns, several of the easiest items were answered incorrectly, whereas some of the most difficult items were answered correctly. The associated small PEs reflect this type of unexpected behavior. For comparison, Table 1 (bottom panel) shows the scores of three persons whose item score patterns were not flagged by the PE statistic. These response patterns are more consistent with the expected behavior: The most difficult items were answered incorrectly more often than were the easy items.

Asymptotic Distribution of the PE Statistic

One limitation of the PE statistic is that its computation requires a complete enumeration of all response patterns with the same length and total-correct score as the response pattern under inspection. This task becomes demanding for numbers of items larger than, say, 20 on an average personal computer (see Table A2 in the Online Appendix for an illustration of how quickly the total number of response vectors increases as the number of items increases). Depending on the number of items, it might be possible to circumvent the problem by using supercomputers. Nevertheless, it would be useful to approximate the exact distribution of the PE statistic through an asymptotic distribution for long tests. In Part H of the Online Appendix, a statistical derivation (based on previous work by van der Flier, 1982) is shown, which was used as an attempt to approximate the exact distribution of PE for large tests. The authors confirmed that this approximate distribution worked well only for a very limited range of situations (i.e., when all the p values are very close to each other) for tests consisting of 20 items. Hence, it is still not clear how many items are required for the approximate distribution to be useful for long tests. More research that can clarify this issue, or that possibly presents different distributional alternatives, is still needed and should be the focus of future research.

Discussion

In this study, the authors discussed a nonparametric statistic to detect misfitting item score patterns that is based on complete enumeration of all possible item score patterns. A big advantage of this method compared with existing methods is that practitioners can use the PE using a prespecified probability level. A drawback is that it can only be used for tests of moderate length due to the rapid increase of computational labor as the number of items increases. It is important to observe that the procedure used here does not guarantee that aberrant behavior did indeed take place whenever a flagging occurs. The PE, as is usually the case for interpreting person-fit statistics, can only provide indications of presence of aberrant behavior. The PE should not be used as conclusive evidence that aberrant behavior *did* occur. Some follow-up strategies (e.g., interviewing the flagged examinees, interviewing the proctors, consulting the seating charts) could provide more substantive information.

The authors applied this method to empirical data measuring different personality traits, for which they first checked the assumptions using a methodology proposed by Sijtsma et al. (2011). An interesting observation is that not all items complied to the MHM model when they used $c = .3$ as a lower bound for item scalability coefficients. This somewhat reduced the number of items that could be used to determine person fit. A researcher finds himself then in the vexing position of having to remove items because of inferior psychometric quality and keeping items in the scale because longer tests are better suited to detect person misfit (Meijer, Sijtsma, & Molenaar, 1995). There are good arguments, however, to first investigate the scale quality of a set of items before conducting person-fit research. An important argument is that inspecting the psychometric quality of the items and removing items with insufficient quality reduces the error component when one tries to interpret misfitting response behavior. When an

item cannot be described by an IRT model (e.g., because it correlates negatively with other items), or when an item has low discrimination (low H_i value), its score is a very unreliable indicator of the latent variable that a researcher is trying to measure. Taking the item scores on these items into account to assess person fit increases the error component in the score patterns, and thus hinders the (psychological) interpretation of these scores. Note that the PE only relates the items with the proportion-correct scores and thus does not account for the discrimination of an item. Thus, it is important to check for items with low discriminating power, as the authors did.

Finally, in this study, the authors discussed the PE for dichotomous items, which is a type of item that is often encountered in educational and intelligence testing. However, this procedure can be generalized to polytomous item scores, which will be a topic for future research.

Authors' Note

The opinions and conclusions contained in this report are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

The online appendix is available at http://apm.sagepub.com/supplemental

Acknowledgments

The authors thank two anonymous reviewers and the editor for their help with improving this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study received funding from the Law School Admission Council.

References

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*, 269-306.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 171-191.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224-247.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, *10*, 101-119.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing*, *10*, 295-317.

Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, *52*, 718-722.

Geisinger, K. F. (2012). Worldwide test reviewing at the beginning of the twenty-first century. *International Journal of Testing*, *12*, 103-107.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139-150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Claussen (Eds.), *Measurement and precision* (pp. 60-90). Princeton, NJ: Princeton University Press.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331-347.

International Test Commission. (2011). *ITC guidelines for quality control in scoring, test analysis, and reporting of test scores*. Available from http://intestcom.org

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277-298.

Lee, P. M. (2004). *Bayesian statistics: An introduction*. West Sussex, England: John Wiley.

Ligtvoet, R., van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*, 578-595.

Luteijn, F., van Dijk, H., & Barelds, D. P. H. (2005). *NPV-J: Junior nederlandse persoonlijkheidsvragenlijst. herziene handleiding 2005* [NPV-J: Dutch personality questionnaire-junior: Professional manual (revised)]. Amsterdam, the Netherlands: Harcourt Assessments.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's lz* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57-81.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*, 311-314.

Meijer, R. R., & Egberink, I. J. L. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement*, *72*, 589-607.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, *90*, 227-238.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*, 261-272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT model. *Applied Psychological Measurement*, *19*(4), 323-335.

Meijer, R. R., & Tendeiro, J. N. (2012). The use of the lz and lz* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, *37*, 758-766.

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2013). *The use of nonparametric IRT to explore data quality*. Manuscript in preparation for the Handbook of Item Response Theory Methods as Applied to Patient Reported Outcomes.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417-430.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for windows*. Groningen, the Netherlands: IEC ProGAMMA.

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79-105.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191-207.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*, 31-37.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331-342.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, *35*, 419-432.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1-19.

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*, 1-27.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267-298.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24*, 65-77.